Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization

Xiaofei Li, Laurent Girin, Radu Horaud, Sharon Gannot

Multiple-Speaker Localization Based on Direct-Path Features and Likelihood Maximization with Spatial Sparsity Regularization

Xiaofei Li, Laurent Girin, Radu Horaud and Sharon Gannot

***Abstract*—This paper addresses the problem of multiple-speaker localization in noisy and reverberant environments, using binaural recordings of an acoustic scene. A complex-valued Gaussian mixture model (CGMM) is adopted, whose components correspond to all the possible candidate source locations defined on a grid. After optimizing the CGMM-based objective function, given an observed set of complex-valued binaural features, both the number of sources and their locations are estimated by selecting the CGMM components with the largest weights. An entropy-based penalty term is added to the likelihood to impose sparsity over the set of CGMM component weights. This favors a small number of detected speakers with respect to the large number of initial candidate source locations. In addition, the direct-path relative transfer function (DP-RTF) is used to build robust binaural features. The DP-RTF, recently proposed for single-source localization, encodes inter-channel information corresponding to the direct-path of sound propagation and is thus robust to reverberations. In this paper, we extend the DP-RTF estimation to the case of multiple sources. In the short-time Fourier transform domain, a consistency test is proposed to check whether a set of consecutive frames is associated to the same source or not. Reliable DP-RTF features are selected from the frames that pass the consistency test to be used for source localization. Experiments carried out using both simulation data and real data recorded with a robotic head confirm the efficiency of the proposed multi-source localization method.**

***Index Terms*—Multiple-speaker localization, candidate-based GMM, entropy penalty, direct-path RTF.**

I. Introduction

Multiple-speaker localization is an auditory scene analysis module with many applications in human-computer and human-robot interaction, video conferencing, etc. In this paper we address the multiple-speaker localization problem in the presence of noise and in reverberant environments. While we use binaural recordings of the acoustic scene, the method can be easily generalized to an arbitrary number of microphones.

Whenever there are more sources than microphones, which is the case in the present work, the so-called W-disjoint orthogonality (WDO) of the speech sources [1], [2] is widely employed by multiple-speaker localization methods. The principle is that in each small region of the time-frequency (TF) domain, the audio signal is assumed to be dominated by only one source, because of the natural sparsity of speech signals in this domain. Therefore, multiple-speaker localization from binaural recordings can be decomposed in the following three-step process: (i) binaural TF-domain localization features are extracted from the binaural signals using the short-time Fourier transform (STFT), or another TF decomposition; (ii) these features are clustered into sources, and (iii) the clustered features are mapped to the source locations.

Traditionally, the binaural features used for localization are the interaural level difference (ILD) and interaural time (or phase) difference (ITD or IPD), e.g., [2], [3], [4], [5], [6]. Complex-valued features can also be used [7], [8], [9], as well as the relative phase ratio [10], [11]. However, these features are not robust to noise and reverberations. To reduce the noise effects, unbiased relative transfer function (RTF) estimators were adopted, such as the ones based on noise stationarity versus the non-stationarity of the desired signal [12], [13], [14], on speech presence probability and spectral subtraction [14], [15], [16], or on complex t-distribution [17]. The RTF estimation is generalized to multiple sources in [18]. To robustly estimate localization features in the presence of reverberations, the precedence effect [19] can be exploited, relying on the principle that signal onsets are dominated by the direct path. In [20], the TF bins dominated by one same source are grouped together based on the use of monaural features (such as pitch and onset/offset). Interaural coherence [21], coherence test [22] and direct-path dominance test [23] were also proposed to detect the frames dominated by one active source. However, in practice, significant reverberations often remain in the selected frames, due to an inaccurate model or to an improper decision threshold. In [24], [25] it was proposed to use direct-path RTF (DP-RTF) binaural features, i.e. the ratio between the direct path of the acoustic transfer function of the left and right channels. Unlike RPR, RTF and similar features, which are polluted by reverberations, the DP-RTF bears mainly the desired localization information. The DP-RTF is estimated based on the convolutive transfer function (CTF) approximation [26], [27] in the STFT domain. The CTF is a convolutive filter on the STFT coefficients of source signal, thus it is a more accurate representation of the STFT-domain binaural signals than the conventional multiplicative transfer function (MTF) approximation [28]. In [24], [25], it

X. Li and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France. E-mail: first.last@inria.fr

L. Girin is with INRIA Grenoble Rhône-Alpes and with Univ. Grenoble Alpes, GIPSA-lab, Grenoble, France. E-mail: laurent.girin@gipsa-lab.grenoble-inp.fr

Sharon Gannot is with Bar Ilan University, Faculty of Engineering, Israel. E-mail: Sharon.Gannot@biu.ac.il

This work was supported by the EU FP7 STREP project EARS #609465 and by the ERC Advanced Grant VHIA #340113.

was observed that single-source localization using the DP-RTF features outperformed MTF-based features.

To localize multiple active speakers using binaural features, many models have been developed. The simplest one, assuming free-field recording with small inter-microphone distance and low reverberations, rely on frequency-independent ITD features. Histogram methods [2], [21] and k-means clustering [7] were then proposed to group these features and localize/separate the sources. When the inter-microphone distance is larger, the problem becomes more complex since the features derived from phase measures (IPD and ITD) are generally ambiguous along frequency due to phase wrapping. In [3], [4], the ITD ambiguity along frequency is solved by jointly exploiting the ILD. Frequency-wise clustering can be adopted, such as hierarchical clustering [8] and weighted sequential clustering [9]. This approach faces the so-called source permutation problem, i.e. the indexing of clusters can be different from one frequency to the other. To solve this problem, the speech spectrum correlation between adjacent frequencies is exploited in [9]. A maximum likelihood method is proposed to formulate the source localization problem in [29]. Based on manifold learning, two semi-supervised localization methods are proposed in [30], [31]. A probabilistic mixture of linear regressions is used in [32] to map a high-dimensional binaural feature vector (concatenated across frequencies) onto source location. In [32], only one source is considered. In [33] the method is extended to multiple sources relying on the WDO assumption. In [34] it is also extended to the direct colocalization of two sources without relying on the WDO assumption and source clustering.

Often, solving the IPD ambiguity and/or source permutation problems amounts to ensure the continuity of binaural cues across frequencies. IPD profiles as a function of frequency can be unwrapped using the direct-path propagation model. In [35], [36], permutation alignment is processed by minimizing the cost function between the observations and the propagation model. In [20], the azimuth set that has the largest likelihood given the feature observations is exhaustively searched from all the potential azimuth sets. Probabilistic models, mostly Gaussian mixture models (GMMs), were also proposed to both cluster and map the features onto source location [3], [5], [6], [10]. In [5] a GMM is used to learn offline the azimuth-dependent ambiguous ITD space of candidate sources. Then, the most likely azimuth with respect to the observed ITDs is estimated as the source direction. In [6] a mixture of warped lines is fitted to the IPD observation profiles. Each warped line corresponds to a source direction. In [3] each candidate interchannel time delay is considered as a GMM component. A mixture of GMMs is constructed to represent multiple sources. The azimuth of each source is given by the component that has the highest weight in the corresponding GMM. A similar approach is proposed in [10], but with GMM components corresponding to candidate 2D source positions thanks to the use of several pairs of microphones.

Recently, a probabilistic clustering method was proposed in [11] to localize an unknown number of emitting speech sources hypothetically located on a regular grid, where each grid point location is known with respect to several microphone pairs. The relative phase ratio (RPR) associated with a microphone pair is predicted from the propagation model for each grid point and for each frequency. A set of complex-valued Gaussian mixture models (CGMMs), one mixture model per frequency, is built such that i) the number of components of each mixture equals the number of grid points, ii) the mixture components are centered around the predicted RPRs (which are frequency-dependent), iii) all mixture components share the same fixed variance which is frequency-independent, and iv) the mixtures at different frequency bins share the same set of weights. Note that unlike [3] and [10] that use a separate GMM for each source, a common CGMM is used for all sources in [11]. Since the mixture means (predicted RPRs) and the variances are fixed, only the mixture weights have to be estimated, one weight corresponding to one candidate source location. An EM algorithm alternates between assigning RPR observations to the mixture components (expectation) and estimating the weights (maximization). At convergence, the algorithm yields a weight value for each grid point and the number and location of active sources is obtained by applying a threshold to these weight values. Because in the present multiple-source localization problem the number of actual sources is expected to be much lower than the number of candidate source locations on the grid, it makes sense to design a methodology to ensure the sparsity of the estimated weights, i.e. to ensure that only a few number of weights have a large value. Obviously, this is expected to facilitate the selection of relevant sources.

This idea is connected to sparse finite mixture modeling, namely to deliberately specify an overfitting mixture model with too many components and looking for sparse solutions in terms of the number of components [37], [38], [39]. Solutions have been proposed in a Bayesian formulation within either variational inference [40] or sampling strategies [39]. To obtain a sparse solution an appropriate prior on the weight distribution must be set and a popular choice is the Dirichlet distribution [39], [40], [41]. The choice of the hyper-parameter of this distribution must guarantee that superfluous components are emptied automatically. While the use of Dirichlet priors is appealing from a Bayesian perspective, several problems appear in practice. First, it is not clear how to learn from the data how much sparsity is needed, i.e. how to choose the hyper-parameter. The rigorous asymptotic analysis of [38] suggests that the Dirichlet hyper-parameter should be smaller than half the dimension of the parameter vector characterizing a mixture component. But in the case of a finite number of observations, a much smaller value seems appropriate [40], [39]. Second, one has to remove the emptied components by checking for small weights, which amounts to thresholding the Dirichlet posterior distribution. Third, in a multi-source localization model such as the one in [11], the means are constrained by the acoustic model, therefore a full Bayesian treatment may not be justified.

Imposing a spatially sparse solution to multi-source localization has also been investigated in [42] in a source signal reconstruction framework. The multiple sources are also

hypothetically located on a regular grid. The mixing matrix is composed of the steering vectors to all grid points, which are given by the free-field sound propagation model under anechoic assumption. The signal reconstruction is formulated as an ℓ_2 fit between the received signal and the mix of source signals. An ℓ_1 regularization is used to impose that only a few grid points correspond to active sources. In [43], this method is extended to the reverberant environment.

In the present paper a new multiple-speaker localization method is proposed, based on the CGMM of [11] and associated likelihood function, combined with the use of direct-path relative transfer function (DP-RTF) as a binaural feature [24]. This paper has the following contributions.

- First, it is proposed to minimize the negative log-likelihood function by adding an entropy-based penalty which enforces a sparse solution in terms of the free model parameters, i.e. the component weights. This corresponds to enforcing the spatial sparsity of sources, in the spirit of [42] but implemented in a very different manner.
- Second, it is shown that the minimization of this penalized objective function can be carried out via a convex-concave optimization procedure (CCP) [44], [45]: at each iteration, the concave penalty is approximated by its first-order Taylor expansion, such that the convex-concave problem becomes convex. The latter is solved using the primal-dual interior point method (PDIPM) [46].
- Third, in the single-speaker configuration of [24], the DP-RTF was estimated at each frequency by solving a unique multi-dimensional linear equation built from the statistics of the binaural signals using all available time frames. However, for multiple sources, successive time frames at a given frequency may not belong anymore to a single source, and one has to enforce the WDO assumption. At each frequency, the multi-dimensional linear equation used for estimating the DP-RTF is now constructed from a frame region (a set of continuous frames) where only one source is assumed to be active.
- Fourth, since the above extension is far from being trivial due to multi-source overlap, a consistency-test algorithm is proposed to verify whether a frame region is associated with a single source or not. If so, a *local* DP-RTF estimation is obtained by solving this local equation, otherwise this frame region is discarded. Applying this principle to many different regions over the entire binaural power spectrogram leads to a set of DP-RTF estimates, each one assumed to correspond to one of the sources.

Overall, these contributions lead to an efficient multiple-source localization method in the presence of noise and reverberations.

The remainder of the paper is organized as follows. Section II provides an overview of the method. The CGMM with maximization of penalized likelihood is described in Section III. The estimation of DP-RTF from the microphone signals for the case of multiple speakers is presented in Section IV. Experiments with both simulated and real data are presented in Section V. Section VI concludes the paper.

II. Method overview

In this section, we briefly specify the articulation of the different processing blocks that have been introduced in the introduction, to provide a clear overview of the proposed method. To this aim, a block diagram of the proposed method is given in Fig. 1. The process in each block will be detailed in the following parts of the paper (the corresponding subsections are specified below).

The microphone signals are first transformed into the STFT domain (Block ①). Based on the consistency test, the DP-RTF features that are respectively associated to one single active speaker are estimated from the STFT coefficients (Block ② and ③). These are detailed in Section IV-B and IV-C. These estimated DP-RTFs are suitable for the CGMM clustering framework presented in Section III: Predicted DP-RTFs (which are the means of the CGMM) are calculated offline from a reverberation-free propagation model. In this work, we use head-related transfer functions (HRTFs) since the recordings are made from either a dummy-head or a robot head (Block ④). Then, the measured and predicted DP-RTF features are provided to the CGMM penalized likelihood maximization procedure (Block ⑤ and ⑥). This process is detailed in Section III-B. This procedure outputs the optimized CGMM component weights for all predefined candidate positions, from which source localization is finally performed using a peak selection routine (Block ⑦). The peak selection procedure depends on the experimental configuration and is thus detailed in Section V.

III. CGMM with Sparsity Regularized Likelihood Maximization

We consider non-stationary source signals $s^i(n)$, e.g. speech, where $i \in [1, I]$ denotes the source index. The received binaural signals are

$$\begin{aligned} \tilde{x}(n) &= x(n) + u(n) = \sum_{i=1}^{I} a^i(n) \star s^i(n) + u(n), \\ \tilde{y}(n) &= y(n) + v(n) = \sum_{i=1}^{I} b^i(n) \star s^i(n) + v(n), \end{aligned} \tag{1}$$

where $x(n)$ and $y(n)$ are the speech mixtures, $u(n)$ and $v(n)$ are the microphone noise signals, $a^i(n)$ and $b^i(n)$ are the binaural room impulse responses (BRIR) from source to microphone, and $\star$ denotes convolution. The binaural signals are transformed into the time-frequency (TF) domain by applying the STFT. As mentioned above, many types of binaural features can be extracted in the TF domain. Let $c_{p,k}$ denote the complex-valued binaural features of interest, where $p \in [1, P]$ is the frame index, and $k \in [0, K-1]$ is the frequency index. The nature of $c_{p,k}$, namely DP-RTF features and their estimation from binaural signals are presented in Section IV, more specifically they are computed in (22). Based on the WDO assumption, a $c_{p,k}$ feature is associated with a single source. However, in practice, some of the TF bins are dominated by noise or by several sources, and hence they should not be considered by the clustering process. Let $\mathcal{P}_k$ denote the set of frame indexes that are associated with a single source at frequency k. Let $\mathcal{C} = \{\{c_{p,k}\}_{p\in\mathcal{P}_k}\}_{k=0}^{K-1}$ denote the

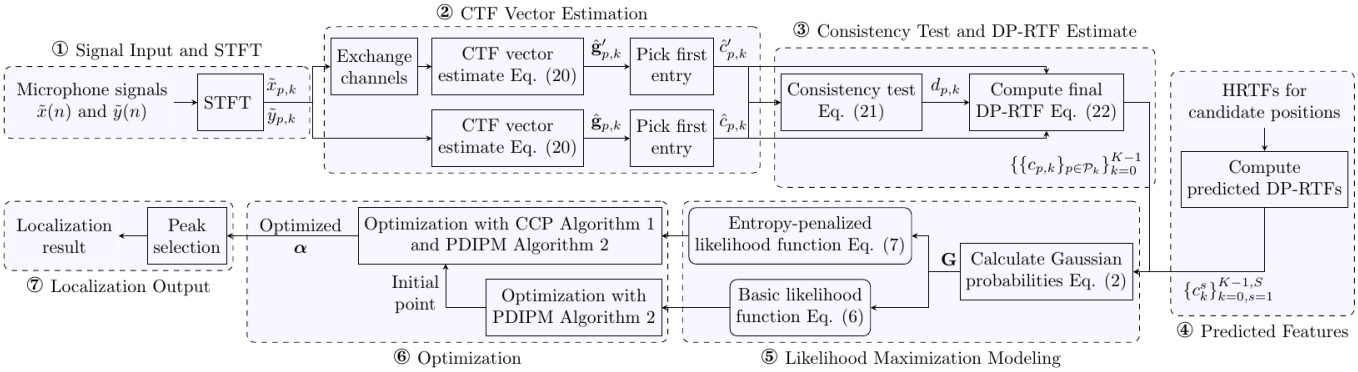


Fig. 1: Flowchart of the proposed sound source localization method.

set of features over all frequencies and available frames. The procedure of selecting "reliable" features (i.e. generating $\mathcal{C}$) will also be detailed in Section IV. In this section, we exploit $\mathcal{C}$ to perform source localization. The multi-source localization problem is cast into a probabilistic clustering problem using a complex-Gaussian mixture model (CGMM).

A. Clustering-Based Localization

In order to group $c_{p,k}$ features into several clusters and hence to achieve multiple-source localization, we adopt the complex-Gaussian mixture model (CGMM) formulation proposed in [11]. Each CGMM component corresponds to a candidate source position on a predefined grid. Source counting and localization are based on the selection of those components having the highest weights. In [11] several pairs of microphones are used so that two-dimensional (2D) localization on a 2D regular grid can be achieved. In this paper, we focus on using a single microphone pair and thus we can only estimate the sources' azimuths [3], [4], [5], [20]. The extension to several microphone pairs is straightforward. We define a set $\mathcal{S}$ of S candidate azimuths regularly placed on a circular grid. In the remainder, $s \in \mathcal{S}$ denotes a candidate azimuth.[1] The probability of an observed binaural feature $c_{p,k} \in \mathbb{C}$, given that it is emitted by a sound source located at s, is assumed to be drawn from a complex-Gaussian distribution with mean $c_k^s \in \mathbb{C}$ and variance $\sigma^2 \in \mathbb{R}$:

$$P(c_{p,k}|s) = \mathcal{N}_c(c_{p,k}; c_k^s, \sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|c_{p,k}-c_k^s|^2}{\sigma^2}\right). \quad (2)$$

The mean c_k^s is the predicted binaural feature at frequency k as provided by a direct-path propagation model. The latter can be derived from the geometric relationship between the microphones and the source candidate position. If an acoustic dummy head is used for the binaural recordings, as will be the case in our experiments, the head-related transfer function (HRTF) of the dummy head is used to predict the means c_k^s by taking the HRTF ratio between channels, for each grid point s and for each frequency k.

[1]For convenience s can indifferently denote a source azimuth or an index of this azimuth within the grid, arbitrarily set from 1 to S.

We now consider the grid of all possible locations, in which case the probability of a binaural feature, given the grid locations, is drawn from a CGMM:

$$P(c_{p,k}|\mathcal{S}) = \sum_{s=1}^{S}\alpha_s\mathcal{N}_c(c_{p,k}; c_k^s, \sigma^2), \quad (3)$$

where $\alpha_s \geq 0$ is the prior probability that the binaural feature is drawn from the s-th component, namely the prior probability that the source is located at s, with $\sum_{s=1}^{S}\alpha_s = 1$. In the present work, α_s is referred to as the component weight. Let us denote the vector of weights with $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_S]^\top$. Since the mixture means are determined based on the source-sensor geometry, and the variance is set to an empirical value σ^2 common to all components and all frequencies,[2] the components of $\boldsymbol{\alpha}$ are the only free model parameters.

Assuming that the observations in $\mathcal{C}$ are independent, the corresponding log-likelihood function (as a function of $\boldsymbol{\alpha}$) is given by:

$$\log\mathcal{L}(\mathcal{C}|\boldsymbol{\alpha}) = \sum_{k=0}^{K-1}\sum_{p\in\mathcal{P}_k}\log\Big(\sum_{s=1}^{S}\alpha_s\mathcal{N}_c(c_{p,k}; c_k^s, \sigma^2)\Big). \quad (4)$$

untitled Multiple-source localization amounts to the maximization of the log-likelihood (4). Importantly, the model above integrates the binaural features of all frequencies by sharing the weights over frequencies, and considers as many components as grid points.[3] Intuitively, after maximization of (4), an active speaker location corresponds to a component with a large weight. In practice, a plot of the weights as a function of azimuth indeed exhibits a quite smooth curve with a few peaks that should correspond to active speakers, see Section V. Therefore, the detection and localization of active speakers could be jointly carried out by selecting the components with the largest weights. A simple strategy would consist of selecting the peaks that are above a threshold, as done in [11], or of selecting the N_s largest peaks if the

[2]This was reported as a relevant choice in [11], and our experiments confirmed that a constant variance outperforms other mechanisms, such as setting the variance to be candidate-dependent (i.e. σ_s^2), or frequency-dependent (σ_k^2), or both ($\sigma_{k,s}^2$).

[3]Note that having one common source location candidate per mixture component shared across frequencies avoid the source permutation problem mentioned in the introduction.

number of active sources N_s is known in advance. However, spurious peaks often appear, due to, e.g., reverberated phantom sources, corrupting the source detection and localization. In the next subsection we propose a penalized maximum likelihood estimator, to enforce a sparse solution for $\boldsymbol{\alpha}$ and remove such spurious peaks.

B. Penalized Maximum Likelihood Estimation

Let $C = |\mathcal{C}|$ denote the cardinality of $\mathcal{C}$, namely the number of binaural observations. We note that (4) can be written as:

$$\log \mathcal{L}(\mathcal{C}|\boldsymbol{\alpha}) = \sum_{c=1}^{C} \log \Big(\sum_{s=1}^{S} g_{cs}\alpha_s \Big) = \mathbf{1}_C^\top \log(\mathbf{G}\boldsymbol{\alpha}), \quad (5)$$

where $\mathbf{1}_C$ denotes a vector in $\mathbb{R}^C$ with all entries set to 1, $\mathbf{G} \in \mathbb{R}^{C\times S}$ is the matrix of probabilities (2) reorganized so that each row $\boldsymbol{g}_c$ of $\mathbf{G}$ corresponds to an observation in $\mathcal{C}$ and each column corresponds to a candidate source position, and where we used the notation:

$$\log(\mathbf{G}\boldsymbol{\alpha}) = \big[\log(\boldsymbol{g}_1\boldsymbol{\alpha}), \ldots, \log(\boldsymbol{g}_c\boldsymbol{\alpha}), \ldots, \log(\boldsymbol{g}_C\boldsymbol{\alpha}) \big]^\top.$$

Then, the maximization of the log-likelihood (4) can be written as the following convex optimization problem:

$$\begin{aligned} \text{minimize} \quad & -\mathbf{1}_C^\top \log(\mathbf{G}\boldsymbol{\alpha}) \\ \text{s.t.} \quad & -\boldsymbol{\alpha} \preceq \mathbf{0}_S, \quad \mathbf{1}_S^\top \boldsymbol{\alpha} = 1, \end{aligned} \quad (6)$$

where $\mathbf{0}_S$ denotes a vector in $\mathbb{R}^S$ with all entries set to zero, and $\preceq$ denotes entry-wise vector inequality. This convex optimization problem with equality and inequality constraints can be solved by the primal-dual interior-point method (PDIPM) [46], which will be described in Section III-C. This optimization problem has the same solution as the original problem of maximizing the log-likelihood (4). However, in the following, we introduce a regularization term to impose the sparsity of $\boldsymbol{\alpha}$, which can be easily added to (6), but cannot be easily added to (4) within an EM algorithm.

We remind that the parameter α_s is the prior probability of having an active source at location s. In practice, the number of active speakers is much lower than the number of candidate locations on the grid. One may consider a grid with tens or hundreds of source locations, but only a handful of this locations correspond to actual sources. Therefore, we may seek a sparse vector $\boldsymbol{\alpha}$ i.e. with only a few nonzero entries. To enforce the sparsity of $\boldsymbol{\alpha}$ we propose to add a penalty term to the objective function in (6). The entries of $\boldsymbol{\alpha}$ are probability masses of a discrete random variable. Generally, the sparser the vector, the smaller entropy $H(\boldsymbol{\alpha}) = -\boldsymbol{\alpha}^\top \log(\boldsymbol{\alpha})$ is. Therefore, the entropy may be used as the required penalty. A sparse solution for $\boldsymbol{\alpha}$ can be obtained by solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{C}\mathbf{1}_C^\top \log(\mathbf{G}\boldsymbol{\alpha}) - \gamma \boldsymbol{\alpha}^\top \log(\boldsymbol{\alpha}) \\ \text{s.t.} \quad & -\boldsymbol{\alpha} \preceq \mathbf{0}_S, \quad \mathbf{1}_S^\top \boldsymbol{\alpha} = 1 \end{aligned} \quad (7)$$

where $\frac{1}{C}$ plays the role of a normalization factor, and γ is an empirical parameter that enables to control the trade-off between the log-likelihood and the entropy.

The entropy $-\boldsymbol{\alpha}^\top \log(\boldsymbol{\alpha})$ is a concave function. Thence the problem can be solved via a convex-concave procedure (CCP) [44]. To solve the CCP, an iterative method is proposed in [45], [47]. At each iteration, the concave function is approximated by its first-order Taylor expansion, so that the convex-concave function becomes a convex function. The derivative of the entropy w.r.t. $\boldsymbol{\alpha}$ is $-(1 + \log(\boldsymbol{\alpha}))$ and the first-order Taylor expansion at $\tilde{\boldsymbol{\alpha}}$ is

$$T_H(\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}}) = -\tilde{\boldsymbol{\alpha}}^\top \log(\tilde{\boldsymbol{\alpha}}) - (\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}})^\top (1 + \log(\tilde{\boldsymbol{\alpha}})).$$

The solution to (7) is summarized in Algorithm 1 which is referred to as EP-MLE (entropy-penalized maximum likelihood estimator). A convergence proof of this procedure is provided in [45], [47]. Subproblem (8) is a convex optimization problem with equality and inequality constraints and, again, it is solved with PDIPM. The algorithm is stopped when the decrease of the objective function (7) from one iteration to the next is lower than a threshold δ. CCP can have (many) local minima, therefore the initialization is important for searching the global minimum, just as for EM algorithms. If γ is small, we assume that the global minimum is in the close proximity of the minimum of (6). Therefore, the initialization of Algorithm 1 is set as the solution of (6), obtained with PDIPM.

Algorithm 1 Concave-convex minimization

Set $m = 0$, initialize $\boldsymbol{\alpha}^{(0)}$ with the solution of (6).
repeat
1 Set $m := m + 1$
2 Solve the convex optimization problem:

$$\begin{aligned} \boldsymbol{\alpha}_{opt} = \underset{\boldsymbol{\alpha}}{\text{argmin}} \{ & -\frac{1}{C}\mathbf{1}_C^\top \log(\mathbf{G}\boldsymbol{\alpha}) + \gamma T_H(\boldsymbol{\alpha}, \boldsymbol{\alpha}^{(m-1)}) \} \\ \text{s.t.} \quad & -\boldsymbol{\alpha} \preceq \mathbf{0}_S, \quad \mathbf{1}_S^\top \boldsymbol{\alpha} = 1 \end{aligned} \quad (8)$$

3 Set $\boldsymbol{\alpha}^{(m)} := \boldsymbol{\alpha}_{opt}$
until Convergence

C. The Primal-Dual Interior-Point Method

We follow [46] to solve for both (6) and (8). [46] provides a general optimization algorithm for a convex objective function f_0 with a set of inequality constraints of the form $f \preceq \mathbf{0}$ and an affine equality constraint. Here $f_0(\boldsymbol{\alpha})$ is the objective function in (6) or (8), and $f(\boldsymbol{\alpha}) = -\boldsymbol{\alpha}$. It is obvious that there exist feasible points for the convex problem (6) and (8), namely the Slater's constraint qualification is satisfied. Therefore, the *strong duality* holds for the present problems, in other words, the optimal duality gap is 0.

PDIPM makes the inequality constraints implicit in the objective function by applying the logarithmic barrier function. As for an inequality constraint $f \leq 0$, the logarithmic barrier

$$\hat{I}\_(f) = -(1/t)\log(-f)$$

is added to the objective. $\hat{I}\_(f)$ takes the value ∞ for $f > 0$ to penalize the objective. The logarithmic barrier is desirable due to its convexity and differentiability. Here t sets the accuracy

of the logarithmic barrier approximation, the larger t, the better the approximation.

The optimization can be expressed as solving the Karush-Kuhn-Tucker (KKT) conditions:

$$r_t(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \nu) = \begin{bmatrix} \nabla f_0(\boldsymbol{\alpha}) - \boldsymbol{\lambda} + \nu \mathbf{1}_S \\ \mathrm{diag}(\boldsymbol{\lambda})\boldsymbol{\alpha} - (1/t)\mathbf{1}_S \\ \mathbf{1}_S^\top \boldsymbol{\alpha} - 1 \end{bmatrix} = \mathbf{0} \tag{9}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^S$ and $\nu \in \mathbb{R}$ are auxiliary variables that originate in the use of the Lagrange multiplier associated with the inequality and equality constraints, respectively. The nonlinear KKT conditions can be solved by Algorithm 2, with the update rule in Step 4 given by the Newton method:

$$\begin{bmatrix} \boldsymbol{\alpha}^{(n+1)} \\ \boldsymbol{\lambda}^{(n+1)} \\ \nu^{(n+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^{(n)} \\ \boldsymbol{\lambda}^{(n)} \\ \nu^{(n)} \end{bmatrix} - \begin{bmatrix} \nabla^2 f_0(\boldsymbol{\alpha}^{(n)}) & -\mathbf{I} & \mathbf{1}_S \\ \mathrm{diag}(\boldsymbol{\lambda}^{(n)}) & \mathrm{diag}(\boldsymbol{\alpha}^{(n)}) & \mathbf{0}_S \\ \mathbf{1}_S^\top & \mathbf{0}_S^\top & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} \nabla f_0(\boldsymbol{\alpha}^{(n)}) - \boldsymbol{\lambda}^{(n)} + \nu \mathbf{1}_S \\ \mathrm{diag}(\boldsymbol{\lambda}^{(n)})\boldsymbol{\alpha}^{(n)} - (1/t^{(n)})\mathbf{1}_S \\ \mathbf{1}_S^\top \boldsymbol{\alpha}^{(n)} - 1 \end{bmatrix} \times \zeta^{(n)} \tag{10}$$

where $^{(n)}$ denotes the iteration index, $\mathbf{I}$ is the identity matrix, and $\zeta^{(n)}$ is the step-length. In the present study, the jth entry of the derivative vector of $f_0(\boldsymbol{\alpha})$ is given by:

$$\nabla f_0(\boldsymbol{\alpha})_j = \begin{cases} -\sum_{i=1}^{C} \frac{g_{ij}}{\sum_{j=1}^{S} g_{ij}\alpha_j}, & \text{for (6)} \\ -\sum_{i=1}^{C} \frac{g_{ij}}{\sum_{j=1}^{S} g_{ij}\alpha_j} - \\ \quad \gamma(1 + \log(\alpha_j^{(m-1)})), & \text{for (8) (at iteration } m) \end{cases} \tag{11}$$

where g_{ij} is the (i,j)-th entry of $\mathbf{G}$. For both (6) and (8), the (j_1, j_2)-th entry of the Hessian matrix is:

$$\nabla^2 f_0(\boldsymbol{\alpha})_{j_1 j_2} = \sum_{i=1}^{C} \frac{g_{ij_1} g_{ij_2}}{(\sum_{j=1}^{S} g_{ij}\alpha_j)^2}. \tag{12}$$

Note that the update rule (10) integrates the fact that the derivative of the inequality function $f(\boldsymbol{\alpha})$ is $\nabla f(\boldsymbol{\alpha}) = -\mathbf{I}$ and that the Hessian matrix of one inequality function $f_s(\boldsymbol{\alpha}) = -\alpha_s$ is $\nabla^2 f_s(\boldsymbol{\alpha}) = \mathbf{0}$ for $s \in [1, S]$.

In Algorithm 2, the primal variable and dual variables are simultaneously updated, and the so-called surrogate duality gap $\hat{\eta}^{(n)}$ is decreasing with the iterations. Correspondingly, the parameter t is increased by the factor μ (a positive value of the order of 10) with respect to $\hat{\eta}^{(n)}$. The line search method for setting the step-length $\zeta^{(n)}$ (Step 3) is briefly summarized in Algorithm 3. Basically, the step-length is set as the largest value that makes the updated variables satisfy the three conditions (i) the dual variable $\boldsymbol{\lambda}$ is nonnegative, (ii) the inequality constraint is satisfied, and (iii) the overall KKT residual is decreased. In this work, the backtracking parameters β and η of Algorithm 3 are set to 0.5 and 0.05, respectively. In the convergence criterion of Algorithm 2, the surrogate duality gap $\hat{\eta}^{(n)}$ is compared with a small threshold ϵ (close to the optimal duality gap, i.e., 0) to guarantee the optimization. The two other criteria are set to guarantee the feasibility of the variables (ϵ_{feas} is also a small arbitrary threshold). For solving (6), a good initialization is to set $\boldsymbol{\alpha}^{(0)} = (1/S)\mathbf{1}_S$, $\boldsymbol{\lambda}^{(0)}$ to an arbitrary positive vector ($10 \cdot \mathbf{1}_S$ in this paper), and $\nu^{(0)}$ to an arbitrary value (0 in this paper). For solving (8) in Algorithm 1, the initialization is set as the solution of the previous iteration. Finally, as already mentioned, Algorithm 1 is initialized by the solution of (6).

Algorithm 2 Primal-dual interior-point

Set $n = 0$, Initialize $-\boldsymbol{\alpha}^{(0)} \preceq \mathbf{0}$, $\boldsymbol{\lambda}^{(0)} \succ \mathbf{0}$, $\nu^{(0)}$.
repeat
1 Compute $\hat{\eta}^{(n)} = \{\boldsymbol{\alpha}^{(n)}\}^\top \boldsymbol{\lambda}^{(n)}$,
2 Set $t^{(n)} := \mu S / \hat{\eta}^{(n)}$,
3 Line search the step-length $\zeta^{(n)}$ (Algorithm 3),
4 Update variables with (10).
until $\hat{\eta}^{(n)} \leq \epsilon$, $\| \mathbf{1}_S^\top \boldsymbol{\alpha}^{(n)} - 1 \|_2 \leq \epsilon_{feas}$, and $\| \nabla f_0(\boldsymbol{\alpha}^{(n)}) - \boldsymbol{\lambda}^{(n)} + \nu \mathbf{1}_S \|_2 \leq \epsilon_{feas}$

Algorithm 3 Line search

Compute $\zeta^{\max} = \sup\{\zeta^{(n)} \in [0,1] | \boldsymbol{\lambda}^{(n+1)} \succeq \mathbf{0}_S\}$, i.e. the largest ζ value that makes the updated $\boldsymbol{\lambda}$ value nonnegative.
Set $\zeta^{(n)} := 0.99\zeta^{\max}$.
repeat
Set $\zeta^{(n)} := \beta\zeta^{(n)}$
until $-\boldsymbol{\alpha}^{(n+1)} \preceq \mathbf{0}_S$ (i.e. the inequality constraint holds) and $\| r_t(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \nu)^{(n+1)} \|_2 \leq (1 - \eta\zeta^{(n)}) \| r_t(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \nu)^{(n)} \|_2$ (i.e. the overall KKT residual is decreased).

IV. Direct-Path Estimation for Multiple Speakers

In this section we propose to estimate the direct-path relative transfer function (DP-RTF) for multiple speakers, which is an extension of the single-speaker case [24]. The rationale of using the DP-RTF is twofold. First, it is robust to noise and reverberations and, second, it is a well-suited binaural feature to be used within the complex-valued generative model (3). For clarity, we first briefly present the single-speaker case [24], and then we move to the multiple-speaker case.

A. DP-RTF Estimation for a Single Speaker

In the case of a single speaker, the noise-free received binaural signals are

$$x(n) = s(n) \star a(n), \quad y(n) = s(n) \star b(n). \tag{13}$$

In the STFT domain, the MTF approximation is only valid when the impulse responses $a(n)$ and $b(n)$ are short, relative to the STFT window. To represent a linear filter with long impulse response in the STFT domain more accurately, the cross-band filters were introduced [26], [48], and a CTF approximation is further introduced and used in [27] to simplify the analysis. Let N and L denote the size and the shift of the STFT window, respectively. Following the CTF, $x(n)$ is approximated in the STFT domain by:

$$x_{p,k} = \sum_{p'=0}^{Q-1} s_{p-p',k} a_{p',k} = s_{p,k} \star a_{p,k}, \tag{14}$$

where $x_{p,k}$ and $s_{p,k}$ are the STFT of $x(n)$ and $s(n)$, respectively, $a_{p,k}$ is the CTF of the filter, and where the convolution $\star$ is executed with respect to the frame index p . The number of CTF coefficients Q is related to the reverberation time. The first CTF coefficient $a_{0,k}$ can be interpreted as the k-th coefficient of the Fourier transform of the impulse response segment $a(n)|_{n=0}^{N-1}$. This holds whatever the actual size of $a(n)$, including if this size is much larger than the STFT window length N. Without loss of generality, we assume that the room impulse response $a(n)$ begins with the impulse response of the direct-path propagation. If the frame length N is properly chosen, $a(n)|_{n=0}^{N-1}$ is thus composed of the direct-path impulse response and possibly of a few reflections. Hence we refer to $a_{0,k}$ as the direct-path acoustic transfer function (ATF). A similar statement holds for $b(n)$ and its corresponding direct-path ATF $b_{0,k}$. By definition, the DP-RTF is given by $\frac{b_{0,k}}{a_{0,k}}$. We remind that the direct-path propagation model in general, and the DP-RTF in particular, have proven to be relevant for sound-source localization.

Based on the cross-relation method [49], using the CTF model of two channels in the noise-free case we have: $x_{p,k} \star b_{p,k} = y_{p,k} \star a_{p,k}$. Dividing both sides by $a_{0,k}$ and reorganizing the terms in vector form we can write:

$$y_{p,k} = \mathbf{z}_{p,k}^\top \, \mathbf{g}_k, \tag{15}$$

where

$$\mathbf{z}_{p,k} = [x_{p,k}, \dots, x_{p-Q+1,k}, y_{p-1,k}, \dots, y_{p-Q+1,k}]^\top$$
$$\mathbf{g}_k = \left[\frac{b_{0,k}}{a_{0,k}}, \dots, \frac{b_{Q-1,k}}{a_{0,k}}, -\frac{a_{1,k}}{a_{0,k}}, \dots, -\frac{a_{Q-1,k}}{a_{0,k}}\right]^\top.$$

We see that the DP-RTF appears as the first entry of the reverberation model $\mathbf{g}_k$. By multiplying both sides of (15) with $y_{p,k}^*$ (the complex conjugate of $y_{p,k}$) and by taking the expectation (in practice averaging the corresponding power spectra over consecutive D frames), we obtain:

$$\hat{\phi}_{yy}(p,k) = \hat{\boldsymbol{\phi}}_{zy}^\top(p,k) \, \mathbf{g}_k, \tag{16}$$

where $\hat{\phi}_{yy}(p,k)$ is the power spectral density (PSD) estimate of $y(n)$ at TF bin (p,k), and $\hat{\boldsymbol{\phi}}_{zy}(p,k)$ is a vector composed of cross-PSD terms between the elements of $\mathbf{z}_{p,k}$ and $y_{p,k}$.

As for the noisy case, an inter-frame spectral subtraction algorithm can be used for noise suppression, e.g. [24]: The auto- and cross-PSD of a frame with low speech power are subtracted from the auto- and cross-PSD of a frame with high speech power. Due to the stationarity of noise and the non-stationarity of speech, the resulting power spectra estimates, $\hat{\phi}_{yy}^s(p,k)$ and $\hat{\boldsymbol{\phi}}_{zy}^s(p,k)$, have low noise power and high speech power. Let $\mathcal{P}_k^s$ be the set of frame indices with high-speech power (at frequency k). After the spectral subtraction, we have:

$$\hat{\phi}_{yy}^s(p,k) = \hat{\boldsymbol{\phi}}_{zy}^s(p,k)^\top \mathbf{g}_k + e(p,k), \quad p \in \mathcal{P}_k^s, \tag{17}$$

with $e(p,k)$ denoting the residual noise of the spectral subtraction procedure. Using the frames indexed in $\mathcal{P}_k^s$, a set of linear equations can be built and solved, yielding an estimate $\hat{\mathbf{g}}_k$ of $\mathbf{g}_k$ and its first component is the estimated DP-RTF.

B. DP-RTF Estimation for Multiple Speakers

As just summarized, all the frames in $\mathcal{P}_k^s$ can be used to construct a DP-RTF estimate in the case of a single speaker. This is no longer valid in the case of multiple speakers, since the frames in $\mathcal{P}_k^s$ do not necessarily correspond to the same source. Hence the DP-RTF estimation method must be reformulated in the case of multiple emitting sources. By applying the STFT to (1), the recorded binaural signals write:

$$\begin{aligned} \tilde{x}_{p,k} &= x_{p,k} + u_{p,k} = \textstyle\sum_{i=1}^I s_{p,k}^i \star a_{p,k}^i + u_{p,k}, \\ \tilde{y}_{p,k} &= y_{p,k} + v_{p,k} = \textstyle\sum_{i=1}^I s_{p,k}^i \star b_{p,k}^i + v_{p,k}. \end{aligned} \tag{18}$$

Without any additional assumption, (17) does not generalize to multiple sources, and thus we cannot directly estimate the DP-RTF associated to each source using the statistics of the mixture signals $x(n)$ and $y(n)$ measured on any arbitrary set of frames. To exploit the above results, we resort to the WDO assumption, i.e. we assume that in a small region of the TF plane only one source is active. Based on this assumption, the DP-RTF in a given TF bin is assumed to correspond to at most one active source. In the following, we thus choose to estimate the DP-RTF for each TF bin. We first formalize this estimate based on the above results. Then we discuss the assumptions for which this estimate is valid and we propose a consistency test to either select the DP-RTF in a given TF bin as a valid estimate for one of the sources or reject it (i.e. we consider that it is not a valid DP-RTF estimate of one of the sources). Using the WDO assumption, and defining:

$$\mathbf{g}_k^i = \left[\frac{b_{0,k}^i}{a_{0,k}^i}, \dots, \frac{b_{Q-1,k}^i}{a_{0,k}^i}, -\frac{a_{1,k}^i}{a_{0,k}^i}, \dots, -\frac{a_{Q-1,k}^i}{a_{0,k}^i}\right]^\top, i \in [1, I],$$

whose first entry is the DP-RTF of source i, we have a possible value $\mathbf{g}_{p,k} \in \{\mathbf{g}_k^i\}_{i=1}^I$ at each STFT bin. In order to estimate $\mathbf{g}_{p,k}$, an equation of the form (17) has to be constructed for a set of frames corresponding to a single source. Let us consider such a set of O consecutive frames to form:

$$\hat{\boldsymbol{\phi}}_{yy}^s(p,k) = \hat{\boldsymbol{\Phi}}_{zy}^s(p,k)\mathbf{g}_{p,k} + \mathbf{e}(p,k), \ \mathbf{g}_{p,k} \in \{\mathbf{g}_k^i\}_{i=1}^I \tag{19}$$

where

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{yy}^s(p,k) &= [\hat{\phi}_{yy}^s(p-O+1,k), \dots, \hat{\phi}_{yy}^s(p,k)]^\top, \\ \hat{\boldsymbol{\Phi}}_{zy}^s(p,k) &= [\hat{\boldsymbol{\phi}}_{zy}^s(p-O+1,k), \dots, \hat{\boldsymbol{\phi}}_{zy}^s(p,k)]^\top, \\ \mathbf{e}(p,k) &= [e(p-O+1,k), \dots, e(p,k)]^\top, \end{aligned}$$

are $O \times 1$ vector, $O \times (2Q-1)$ matrix and $O \times 1$ vector, respectively. Note that (most of) the frames involved in the construction of $\hat{\boldsymbol{\phi}}_{yy}^s(p,k)$ and $\hat{\boldsymbol{\Phi}}_{zy}^s(p,k)$ should have high-speech power, i.e. $[p-O, p] \subseteq \mathcal{P}_k^s$. Assume that $\mathbf{e}(p,k)$ is stationary and independent along frames. Then if the matrix $\hat{\boldsymbol{\Phi}}_{zy}^s(p,k)$ is not underdetermined, i.e. $O \geq 2Q-1$, an optimal estimation of $\mathbf{g}_{p,k}$ is given by the least square solution of (19):

$$\hat{\mathbf{g}}_{p,k} = (\hat{\boldsymbol{\Phi}}_{zy}^s(p,k)^H \hat{\boldsymbol{\Phi}}_{zy}^s(p,k))^{-1} \hat{\boldsymbol{\Phi}}_{zy}^s(p,k)^H \hat{\boldsymbol{\phi}}_{yy}^s(p,k). \tag{20}$$

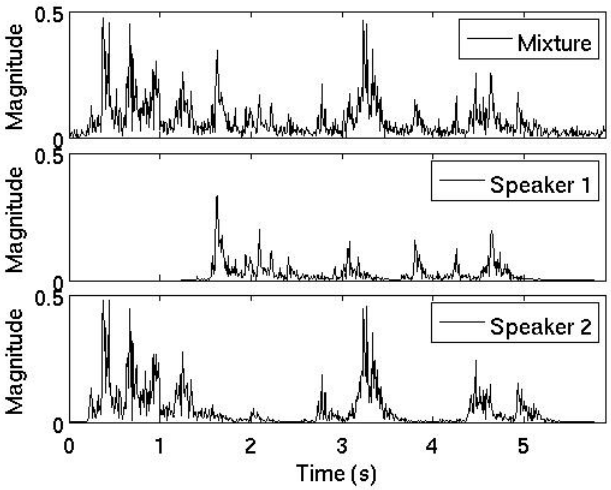
Let σ_k^2 denote the variance of the residual noise $e(p,k)$. The covariance matrix of $\hat{\mathbf{g}}_{p,k}$ is $\sigma_k^2(\hat{\boldsymbol{\Phi}}_{zy}^s(p,k)^H \hat{\boldsymbol{\Phi}}_{zy}^s(p,k))^{-1}$ [50],

which obviously can be reduced by enlarging the number of equations, i.e. O.

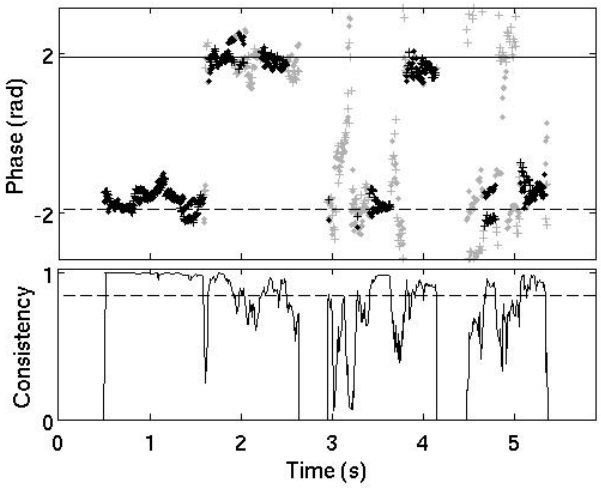
To estimate the cross-PSD between $y_{p-Q+1,k}$ (or $x_{p-Q+1,k}$) and $y_{p,k}$, the past $D-1$ frames before the $(p-Q+1)$-th frame are employed. Therefore, the STFT coefficients in the frame range $[p-Q-D+2,\ p]$ should be associated with a single active speaker. When considering O consecutive frames, the past $Q+D-2$ frames before the $(p-O+1)$-th frame are employed to construct the earliest cross-PSD vector in (19), i.e. $\hat{\boldsymbol{\phi}}^{s}_{zy}(p-O+1,k)$. Therefore, for a correct estimation of the DP-RTF at TF bin (p,k), the STFT coefficients at frequency k in the frame range $[p-O-Q-D+3,\ p]$ should be associated with a single active speaker. In contrast, if the $O+Q+D-2$ consecutive speech frames used in the estimation of a DP-RTF at TF bin (p,k) are composed of coefficients involving multiple active speakers, (20) will not deliver a valid estimate of the DP-RTF, i.e. a DP-RTF estimate that corresponds to one and only one of the sources. In other words, the present work requires a stricter WDO assumption than the original one [1], [2], since at a given frequency bin k, we seek multiple continuous frames associated to a same single source.

In a scenario with multiple and simultaneous speech sources, the natural sparsity and the harmonic nature of speech spectra in the STFT domain make it common that at a given frequency a set of consecutive speech frames is dominated by a single active speaker. However, the amount of speech regions dominated by a single speaker is decreasing with an increasing number of sources I and an increasing CTF length Q. Fig. 2a shows an example of two-speaker mixture at one given frequency (for instance 2 kHz). It can be seen that the magnitude spectrum (at the selected frequency) of individual speech signals exhibits regions with large energy over numerous consecutive frames. This is expected to correspond to a signal harmonic. We observe that most regions are dominated by a single speaker, and as a result, the trajectory of the mixture magnitude coefficient ressembles the sum of the magnitude of the two individual speech signals. This indicates that the WDO assumption can be relaxed to a few hundreds of milliseconds. Note that this mixture is just an example at one given frequency. The overlap between different sources could be much more (or less) than this illustrated example. Taking all the frequency bands into account, there exist a notable number of speech regions dominated by a single speaker for cases where i) the reverberation time is not very long, e.g. not longer than 0.7 s and ii) the number of sources is not very large, e.g. not larger than three.[4] For source localization, we only need a certain number of speech regions to be valid, rather than requiring most of the TF-bins to be valid, as is the case for binary-mask source separation. In the next subsection, we propose a consistency test method to efficiently pick out the valid speech regions.

[4]The requirement of O consecutive frames is to guarantee the least square problem (19) to be not underdetermined. Based on the analysis of the cross-relation method in [49], for the multichannel case, O would be proportional to $\frac{1}{I-1}$. Therefore, O could be reduced, namely WDO assumption could be relaxed, by increasing the number of channels. The case $I>2$ is beyond the scope of this work, and will be investigated in future work.



(a) Magnitude of STFT coefficient



(b) Phase of DP-RTF estimate and consistency test

Fig. 2: An example of multispeaker DP-RTF estimate at a given frequency (2 kHz). Binaural simulations with two speakers at -40° and 40°, SNR = 20 dB, reverberation time $=$ 0.6 s (see the detailed dataset description in Section V-A4). Speaker 1 is active from $\approx$1.5 s to ≈ 5 s, while Speaker 2 is active all the time. (a): Magnitude of STFT coefficient vs. time. (b): Phase of DP-RTF estimates (top) and consistency test (bottom), both on the mixture signal. In the top figure, the dots represent the phase of the DP-RTF estimate, i.e. $\arg[\hat{c}_{p,k}]$, and the cross points (+) represent the estimates after exchanging channels, i.e. $\arg[1/\hat{c}'_{p,k}]$. The markers in black and grey indicate that the TF bins pass the consistency test or not, respectively. The solid line and dashed line denote the predicted phase computed from the HRTFs of Speaker 1 and Speaker 2, respectively. In the bottom figure, the solid curve represents the similarity measure in (21), the dashed line is the threshold set to 0.85. Note that a zero similarity means that there is not enough frames with high speech power in the corresponding region to construct (19).

C. Consistency Test

A consistency test is proposed to check whether a continuous set of $(O+Q+D-2)$ STFT coefficients at a given frequency k are associated with a single active speaker or not. The principle is based on exchanging the roles of the two channels, since the DP-RTF between $b(n)$ and $a(n)$ is the

inverse of the DP-RTF between $a(n)$ and $b(n)$. We thus define $\mathbf{g}'_{p,k}$ as the reverberation model that exchanges the roles of $a_{p,k}$ and $b_{p,k}$ in $\mathbf{g}_{p,k}$. If the STFT coefficients used to estimate $\hat{\mathbf{g}}_{p,k}$ and $\hat{\mathbf{g}}'_{p,k}$ are associated with a single speaker, (15) holds and the two corresponding DP-RTF estimates should be consistent. Conversely, if the STFT coefficients are associated with more than one speaker, or only with reverberations, the estimations $\mathbf{g}_{p,k}$ and $\mathbf{g}'_{p,k}$ are both biased, with inconsistent bias values. As a result, we should observe a discrepancy between the two estimated DP-RTF values.

In practice, let us denote by $\hat{c}_{p,k}$ and $\hat{c}'_{p,k} \in \mathbb{C}$ the first entry of $\hat{\mathbf{g}}_{p,k}$ and of $\hat{\mathbf{g}}'_{p,k}$ respectively, i.e. the DP-RTF estimates $\frac{b_{0,k}}{a_{0,k}}$ and $\frac{a_{0,k}}{b_{0,k}}$. We test the consistency by measuring the difference between $\hat{c}_{p,k}$ and $1/\hat{c}'_{p,k}$. To achieve a normalized difference measurement that allows us to easily set a reasonable test threshold, we define the vectors $\mathbf{c}_{1,p,k} = [1, \hat{c}_{p,k}]^\top$ and $\mathbf{c}_{2,p,k} = [1, 1/\hat{c}'_{p,k}]^\top$, where the first entry 1 can be interpreted as the DP-RTF corresponding to $\frac{a_{0,k}}{a_{0,k}}$. The similarity, i.e. the cosine of the angle, of the two vectors:

$$d_{p,k} = \frac{|\mathbf{c}_{1,p,k}^\top \mathbf{c}_{2,p,k}|}{\sqrt{\mathbf{c}_{1,p,k}^\top \mathbf{c}_{1,p,k} \mathbf{c}_{2,p,k}^\top \mathbf{c}_{2,p,k}}} \tag{21}$$

is a value in $[0, 1]$, which is a good difference measurement. The larger $d_{p,k}$, the more consistent the reverberation model is. The consistency decision is made by comparing $d_{p,k}$ with a threshold d_T (e.g. set to 0.85).

An example of consistency test is shown in Fig. 2b. The test is applied to the mixture signal in Fig. 2a. It can be seen that the phase of $\hat{c}_{p,k}$ and $1/\hat{c}'_{p,k}$ are close to each other, and are close to the predicted phase, for the frames dominated by a single speaker, e.g. within 0.3–1.5 s for Speaker 1. Correspondingly, the consistency measures are large (close to 1). For the regions that involve the two speakers, e.g. around 3.2 s, and the regions that mainly involve the reverberations, e.g. around 3.7 s, the two phase measures are far from the predicted phase, and are far from each other, thus the consistency measures are low. Eventually, the DP-RTF estimates that pass the consistency test are correctly selected, as shown by the black markers, which are close to the predicted value.

Let $\mathcal{P}_k$ denote the set of frames indices that pass the consistency test for frequency k. Every DP-RTF estimation in $\mathcal{P}_k$ is first recalculated as $(\hat{c}_{p,k} + 1/\hat{c}'_{p,k})/2$ to improve the estimate robustness. Finally it is normalized as

$$c_{p,k} = \frac{(\hat{c}_{p,k} + 1/\hat{c}'_{p,k})/2}{1 + |(\hat{c}_{p,k} + 1/\hat{c}'_{p,k})/2|}, \tag{22}$$

which is a complex number whose module is in the interval $[0, 1]$. Each $c_{p,k}$ is assumed to be associated with a single speaker. We thus now have a set of normalized DP-RTF observations that are ready for clustering among sources.

V. Experiments

In this section, we present a series of experiments with simulated data and real data collected from a robotic head. We start by describing the experimental setup, and then give the experimental results and discussions.

A. Experimental Setup

1) Blind and Semi-Blind Configurations: Two configurations were tested, blind and semi-blind. In the blind configuration, the number of active sources I and their locations are simultaneously estimated. Note that the term 'blind' mainly refers to the unknown number of sources, and does not mean a complete blind configuration, for instance the HRTFs and reverberation time are known. Localization is conducted by selecting the local maxima in the set of CGMM weights that are above a threshold α_T, i.e. we detect $\{\alpha | \alpha > \alpha_T, \alpha \in [\alpha_1, \ldots, \alpha_S]\}$. In the semi-blind configuration, I is assumed to be known and the source locations are detected by selecting the I largest local maxima over the weights $[\alpha_1, \ldots, \alpha_S]$. The source location estimates are associated to the ground-truth source locations by looking for the correspondence that provides the overall lower mean absolute localization error (MAE) averaged across sources. In general, blind localization is more difficult than semi-blind localization in terms of peak selection.

2) Performance Metrics: For both configurations, a source is then considered to be successfully localized if the difference between its actual azimuth and the estimated azimuth is not larger than a predefined threshold, empirically set to 15°. Then, a new MAE is calculated for the successfully localized sources, which is the MAE in the results reported below. To further characterize the unsuccessful localizations in the blind configuration scenario, we also calculated: (i) the missed detection (MD) rate defined as the percentage of sources that are present but not detected out of the total number of present sources; and (ii) the false alarm (FA) rate defined as the percentage of sources that are detected although they are not actually present in the scene, out of the total number of sources. In the semi-blind configuration, we calculated the outlier rate, defined as the percentage of sources for which the azimuth error is larger than 15° out of the total number of present sources (in short, the percentage of unsuccessfully localized sources). Note that, on one hand, an outlier indicates a missed detection of the corresponding true source, on the other hand, the outlier estimate itself is a false alarm.

3) Parameters Setting: The signal sampling rate is 16 kHz. Only the frequency band from 0 to 4 kHz is considered for speech source localization, since this band concentrates the largest part of speech signals energy. The setting of the three parameters N, Q and D is crucial for a good estimation of the DP-RTF, and is discussed in [24]. In this work, we use the same parameter setting as in [24], which achieves a good trade-off for various acoustic conditions. The STFT frame length is set to $N = 16$ ms (256 samples) with frame shift $L = 8$ ms (128 samples). The CTF length Q is set to correspond to $T_{60}/6$. The number of frames for the PSD estimate is $D = 15$ (120 ms). We set $O = 3.5Q$ as a trade-off for ensuring a small variance of $\hat{\mathbf{g}}_{p,k}$, and the sparsity of

the speech spectrum (one single active source) on a reasonable number of successive frames. The threshold for the consistency test is set to $d_T = 0.85$. The penalty factor γ in (7) is set to 0.2 as a good experimental trade-off between the log-likelihood and the entropy. The positive factor μ in Algorithm 2 is set to 20. The thresholds for the convergence criterion in Algorithms 1 and 2 are set to $\delta = 10^{-3}$ and $\epsilon = \epsilon_{feas} = 10^{-6}$. In the blind localization configuration, the threshold α_T for the local maximum selection corresponding to source detection is set to 0.05, since this value was shown to provide a good trade-off between MD and FA.

4) Simulated Binaural Data: A set of BRIRs were generated with the ROOMSIM simulator [51] combined with the head-related impulse responses (HRIRs) of the KEMAR dummy head [52]. For the KEMAR dummy head, the pre-measured HRIRs for a large set of discrete directions (for both azimuth and elevation) are available. To simulate the filtering of a reflection coming from a given direction, the pre-measured HRIR of discrete direction closest to the reflection direction is used as an approximation of the HRIR of the reflection direction. This procedure is automatically conducted in the ROOMSIM simulator [51]. The simulated room is of dimension 5 m $\times$ 8 m $\times$ 3 m. The dummy head is located at (1 m, 4 m, 1.5 m). Sound sources were placed in front of the dummy head with azimuths (relative to the dummy head center) varying from $-90°$ to $90°$, spaced by $5°$ (hence 37 azimuths), and an elevation of $0°$. Five sets of 37 binaural signals were generated by selecting 5 different speech signals from the TIMIT dataset [53] and convolving each of these 5 signals with each of the 37 BRIRs.

We set the reverberation time to $T_{60} = 0.6$ s, which is quite notable. Accordingly, we set $Q = 12$ (96 ms) and $O = 42$. Two dummy-head-to-source distances were simulated, namely 1 m and 2 m, for which the direct-to-reverberant ratio (DRR) is about 0.5 dB and -5.5 dB, respectively. Localization of two and three speakers is considered. We generated 500 mixtures for each case, by summing binaural signals randomly selected from the five groups, ensuring that the source directions are spaced by at least $15°$. The noise signals were generated by mixing two types of noise with the same power: (i) directional noise: white Gaussian noise emitted from the source point with azimuth of $120°$, elevation of $30°$ and distance-to-head of 2.2 m, and (ii) spatially uncorrelated white Gaussian noise. The composite noise signal was added to the speech mixture signals with signal-to-noise ratio (SNR) of either 30 dB or 5 dB. The duration of each noisy speech mixture used for localization is of about 3 s. Importantly, in these simulations, the predicted DP-RTF corresponding to the candidate source locations, i.e. the means of the CGMM components, are computed by using the anechoic HRIRs from [52], which ideally corresponds to the direct-path of the complete simulated propagation model (the BRIRs). The set of candidate locations $\mathcal{S}$ is composed of the 37 azimuth values within $[-90°, 90°]$ taken every $5°$.

5) Robotic Head Data: We also report real-world experiments conducted using the head of the NAO humanoid robot (version 5), equipped with four nearly-coplanar microphones,

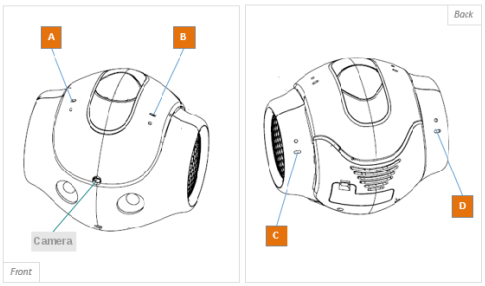


Fig. 3: The four-microphone robot head used in this paper.

see Fig. 3. Elevation localization is here unreliable due to the coplanar microphone array. We used the two microphone pairs A-C and B-D to localize the azimuth relative to the NAO head. The head has built-in fans nearby the microphones, hence the recorded data contain a notable amount of fan noise (aka ego-noise), which is stationary and spatially correlated [54].

The data are recorded in an office room with $T_{60} = 0.52$ s. Accordingly, we set here $Q = 11$ (88 ms) and $O = 38$. The test dataset consists of long speech utterances (> 3 s) from the TIMIT dataset, emitted by a loudspeaker. Two data sets are recorded with a robot-to-source distance of 1.5 m and 2.5 m, respectively (remember that DRR is related to the microphone-to-source distance). For each data set, 174 speech utterances were emitted from directions uniformly distributed in the range $[-120°, 120°]$ for azimuth, and $[-15°, 25°]$ for elevation. The noise of recorded signals mainly corresponds to fan noise, the SNR is about 10 dB. Two-speaker localization and three-speaker localization were considered. For each case, 200 mixtures were generated by summing the sensor signals from two or three different directions. Note that this mixing procedure sums the noise signals from each individual recording, which is different from what would be obtained with a real mixture recording. The summed noise has statistical property identical to the individual noises since latter are identically distributed and stationary, while the SNR is decreased. The mixture signals were truncated to have a duration of 3 s. The source azimuths are spaced by a random angle not lower than $15°$. The candidate azimuths $\mathcal{S}$ are here set to values within $[-120°, 120°]$ with a $6°$-step, hence there are 41 candidate azimuths. As for the two microphone pairs, the predicted binaural features (CGMM mean) of the candidate azimuths were respectively computed by using the corresponding anechoic HRTFs. The HRTFs and the predicted features are computed offline from HRIRs measured in laboratory: white Gaussian noise is emitted from a loudspeaker placed around NAO's head from each candidate direction, and the cross-correlation between the microphone and source signals yields the BRIR of each direction. In order to obtain anechoic HRIRs, the BRIRs are manually truncated before the first reflection.

The information from the two microphone pairs was integrated into the localization model with the following procedure: 1) binaural features are extracted independently from each of the two pairs, 2) the Gaussian probabilities of the binaural features are computed using (2) for each pair; note that the CGMM means c_k^s are different for the two pairs,

but the weights $\boldsymbol{\alpha}$ are of course the same, 3) we have an additional summation over the two pairs of features in the likelihood function (4); this corresponds to have the Gaussian probabilities of the two pairs concatenated into a common matrix $\mathbf{G}$, and finally 4) execute the optimization procedure.

B. Baseline Methods

The results of the proposed method are compared with the results obtained with the four following baseline methods.

1) Basic-CGMM: To test the relevance and efficiency of the entropy penalty, the results obtained with the same CGMM model, but solving the basic optimization problem (6), i.e. without the entropy penalty, are compared. The same proposed DP-RTF feature is used here, and the peak counting threshold of the blind configuration is empirically set to 0.15 to adjust the trade-off between MD and FA.

2) RTF-CT-CGMM: To test the efficiency of the proposed DP-RTF feature, the binaural RTFs with normalized amplitude of [22] are tested for comparison. Here, a coherence test is used to search the TF bins which are supposed to be dominated by one active source. Note that the direct-path source and its reflections are considered as different sources, thence, the TF bins that pass the coherence test are supposed to be dominated by the direct-path signal of one active source. The TF bins that have a coherence larger than a threshold (here set to 0.9) are selected to provide RTF features. The proposed CGMM localization model is used. For the blind configuration, the peak counting threshold is set to 0.15 as a good trade-off between MD and FA. Note that, only the TF bins that have a high speech power are considered for the coherence test. The inter-frame spectral subtraction is applied to the TF bins that pass the coherence test. Therefore the selected RTF features are supposed to have the same robustness to noise as the proposed DP-RTF features.

3) The Model-based EM Source Separation and Localization method (MESSL) [3]: This method is based on a GMM-like joint model of ILD and IPD distribution. MESSL is a semi-blind method, i.e. the number of speakers on a given analyzed sound sequence is assumed to be known. We used the implementation provided by the authors.[5] The default setup is used for the parameter initialization and tying scheme, namely the GMM weights are initialized using a cross-correlation method while the other parameters are initialized in a non-informative way, and the parameters are not tied at all. A pilot comparison was conducted to test the three different configurations: i) default, with ILD but not garbage source, ii) without both ILD and garbage source, and iii) with both ILD and garbage source. The third configuration slightly outperformed the other two, thus it was adopted in the following experiments. For the binaural dataset, the set of candidate delays corresponds to the azimuth grid used for the proposed method, and they are computed from the corresponding HRIRs. For the multichannel robotic head data, the multichannel MESSL proposed in [55] is used. The set of candidate delays is uniformly distributed in the possible maximum range. Source localization is made by comparing the output multichannel delays and the delay templates corresponding to the azimuth grid used for the proposed method.

[5]https://github.com/mim/messl

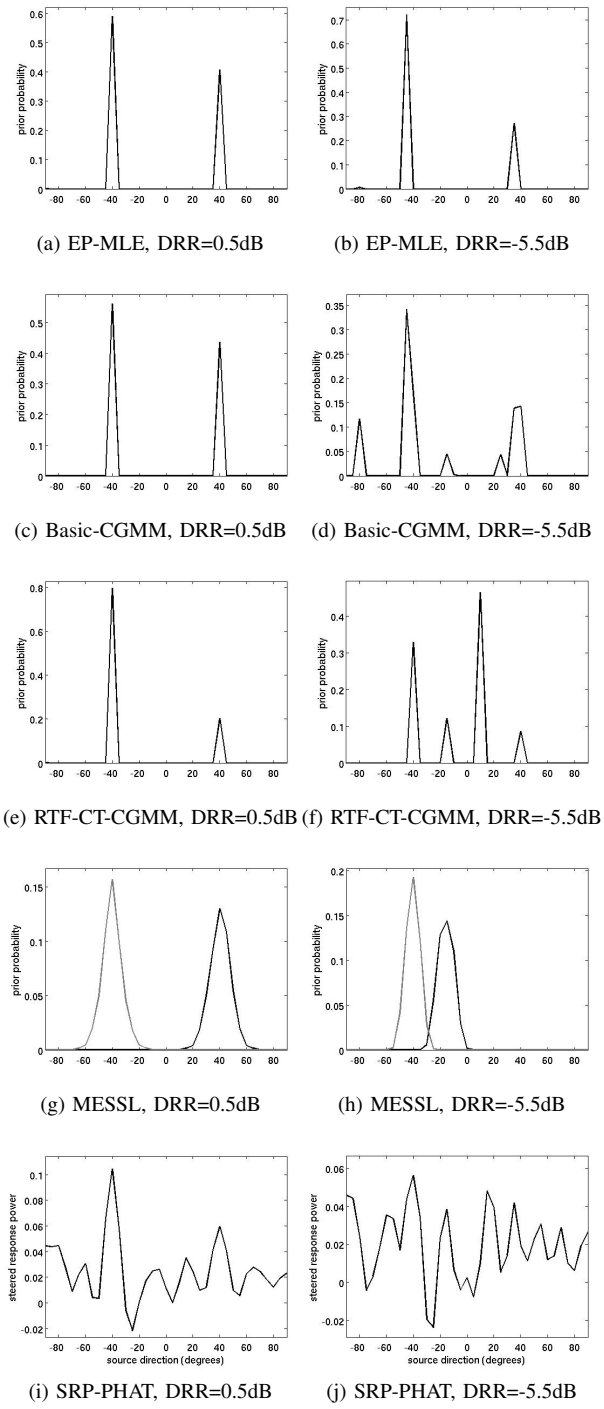
4) The Steered-Response Power using the PHAse Transform (SRP-PHAT) [56], [57]: This is a classic one-stage algorithm. The candidate azimuth directions of the proposed method are taken as the steering directions, and the corresponding HRIRs are used as the steering responses. The number of sources and their locations can be detected by selecting the peaks with steered response power above a threshold. However, the steered response power for different acoustic conditions, such as different number of sources, SNRs, or reverberation times, can significantly vary, which makes the threshold setting difficult. Thence, in the following experiments, we use SRP-PHAT in a semi-blind mode.

C. Results of Experiments with Simulated Data

In this subsection, we first present an example of result obtained on simulated data to illustrate the behavior of the localization methods, and then we provide more general quantitative results. We remind that the proposed method is referred to as EP-MLE.

1) An Example of Sound Source Localization: Fig. 4 shows a source localization example obtained with the proposed method and with the baseline methods. For DRR = 0.5 dB (left column), all methods (except for SRP-PHAT) have two (and only two) prominent peaks at the correct source azimuths. The SRP-PHAT profile is more cluttered than the other profiles but the two highest peaks are nevertheless at the correct source azimuth. The results for the proposed EP-MLE and the Basic-CGMM method are quite similar, hence the entropy penalty has no significant influence in these conditions.

For DRR = -5.5 dB (right column), the source at -40° still has a prominent peak for the first four methods (though the maximum of the peak is slightly shifted at -45° for EP-MLE and Basic-CGMM). Even the SRP-PHAT profile, though made very cluttered by the intense reverberations, keeps its maximum at -40°. However, the source at 40° does not have a very large peak for RTF-CT-CGMM, whereas there is a much higher peak at 10°. One possible reason for this is that a high amount of reverberations decreases the number of TF bins dominated by the direct-path propagation of a single source, hence a lower number of TF bins can be selected by the coherence test. In addition, an improper threshold can make the detected TF bins involve reflections. MESSL fails to detect the source at 40° as well: there are still two prominent peaks but the second one is clearly misslocated at -15°. The reason for this is that the ILD/IPD features are heavily contaminated by strong reverberations. Finally, the very cluttered profile of SRP-PHAT does not allow the detection of the second source. In contrast, it can be seen that the proposed EP-MLE and Basic-CGMM methods provide second-prominent peaks at the correct source location (actually at 35° for EP-MLE).



(a) EP-MLE, DRR=0.5dB (b) EP-MLE, DRR=-5.5dB

(c) Basic-CGMM, DRR=0.5dB (d) Basic-CGMM, DRR=-5.5dB

(e) RTF-CT-CGMM, DRR=0.5dB (f) RTF-CT-CGMM, DRR=-5.5dB

(g) MESSL, DRR=0.5dB (h) MESSL, DRR=-5.5dB

(i) SRP-PHAT, DRR=0.5dB (j) SRP-PHAT, DRR=-5.5dB

Fig. 4: An example of source localization obtained with the proposed method and with the four baseline methods. Two speakers located at azimuths -40° and 40°. SNR is 30 dB.

This again shows that, compared with the MTF-based RTF feature, the proposed DP-RTF feature is more reliable for multi-source localization in highly reverberant environments. In addition to the peaks at the correct azimuths, there are also a few other spurious peaks in the case of the Basic-CGMM method. The use of the entropy penalty in EP-MLE successfully suppresses the spurious peaks and strengthen the true peaks. This illustrates well the sparsity-enforcing property of the entropy penalty term. For the semi-blind configuration, correct localization is obtained by both EP-MLE and Basic-CGMM, in this example. But in the blind configuration, the selection threshold is very difficult to set automatically for the Basic-CGMM method, due to amplitude similarity of the correct peak at 40° and of the spurious peak at -80°. This may easily lead to either miss detection or false alarm. In contrast, the EP-MLE method enables a large range of threshold values that lead to correct detection in this example. Note that there is a larger risk of errors for Basic-CGMM even in the semi-blind configuration: a slightly larger spurious peak at -80° would lead to a wrong localization.

2) Semi-blind Localization Results: Table I shows the semi-blind localization results obtained for various acoustic conditions, averaged over the 500 above-mentioned test mixtures. We first compare the two-speaker localization results of the proposed method with the results of MESSL and SRP-PHAT. For SNR $=$ 30 dB and DRR $=$ 0.5 dB, all three methods achieve satisfactory and comparable performance. When only the DRR decreases (to -5.5 dB), the outlier rate of MESSL and SRP-PHAT dramatically increases, whereas the outlier rate of the proposed method increases only slightly. This indicates that the ILD/ITD features and the steered response power are less robust to reverberations than the proposed DP-RTF features. For MESSL, the garbage source is not able to collect the colored interfering features caused by the intense reverberations. When only the SNR decreases (to 5 dB), the performance measures of all the three methods degrade, as expected. For EP-MLE, the noise residual after spectral subtraction is larger for the low SNR case. Moreover, more frames with low speech power are highly corrupted by noise, which decreases the number of valid TF bins used for DP-RTF estimation. For MESSL, the estimated ILD/ITD features are severely corrupted by the noise, especially by the directional (spatially correlated) noise. In addition, the ILD/ITD extracted from the TF bins dominated by the directional noise will lead to a spurious peak in the noise direction. For these reasons, MESSL performs the worst out of the three methods (at SNR = 5 dB). For SRP-PHAT, the directional noise also contaminates the steered response power, possibly leading to a spurious peak. SRP-PHAT outperforms MESSL, and is comparable with the proposed method, possibly due to the efficiency of PHAT weight. When both SNR and DRR are low (5 dB and -5.5 dB, respectively), the proposed method prominently outperforms the two other methods in terms of outlier rate.

We then analyze the three-speaker localization results. Compared to the two-speaker case, the localization performances of all methods degrade, as expected. Indeed, the WDO assumption is less valid as the number of sources increases, i.e. the number of TF regions that are dominated by a single source decreases. For the proposed method, this leads to a lower number of DP-RTF observations and worse localization performance. For MESSL, this leads to estimated ILD/ITD features that are less reliable, which also leads to a worse localization performance. For SRP-PHAT, the multiple sources can be mutually considered as noise signals, so more sources will make the steered response power of the actual source

TABLE I: Semi-blind localization results for simulation data under various acoustic conditions. The lowest outlier rate among five methods for each acoustic condition is shown in **bold**.

	SNR	DRR	EP-MLE (prop.)		Basic-CGMM		RTF-CT-CGMM		MESSL [3]		SRP-PHAT [57]	
	(dB)	(dB)	Out(%)	MAE(°)	Out(%)	MAE(°)	Out(%)	MAE(°)	Out(%)	MAE(°)	Out(%)	MAE(°)
Two speakers	30	0.5	0.9	0.15	**0.2**	0.18	5.6	1.91	0.4	0.14	2.0	0.42
	30	-5.5	**2.3**	2.06	3.6	2.03	26.4	4.71	27.7	2.06	34.8	2.81
	5	0.5	6.2	1.94	**5.4**	1.94	11.5	4.53	17.4	2.75	6.1	1.75
	5	-5.5	**15.1**	5.12	18.9	5.05	30.1	6.31	36.8	5.13	35.7	4.30
Three speakers	30	0.5	3.4	0.58	**1.5**	0.64	15.5	2.76	2.1	0.46	5.5	0.98
	30	-5.5	**12.9**	2.93	16.1	2.91	29.9	5.54	35.7	2.55	35.6	3.18
	5	0.5	18.7	3.05	17.2	3.08	19.7	5.29	24.1	3.29	**13.4**	2.52
	5	-5.5	**23.1**	5.53	25.6	5.49	33.7	6.64	37.5	5.10	34.7	5.03

TABLE II: Blind localization results for simulation data under various acoustic conditions. The lowest MD and FA among three methods for each acoustic condition are shown in **bold**.

	SNR	DRR	EP-MLE (prop.)			Basic-CGMM			RTF-CT-CGMM		
	(dB)	(dB)	MD(%)	FA(%)	MAE(°)	MD(%)	FA(%)	MAE(°)	MD(%)	FA(%)	MAE(°)
Two speakers	30	0.5	6.2	**0**	0.15	**1.8**	1.5	0.17	11.9	12.0	1.81
	30	-5.5	**4.1**	**6.6**	1.75	9.1	6.7	1.75	28.3	37.7	5.03
	5	0.5	**13.4**	**0.3**	1.68	17.4	1.2	1.70	14.4	17.3	4.45
	5	-5.5	**16.1**	**15.7**	4.88	21.7	17.3	4.79	30.5	37.4	6.68
Three speakers	30	0.5	**17.9**	**0.2**	0.53	18.5	0.5	0.48	27.1	10.0	2.57
	30	-5.5	**19.9**	**9.3**	2.61	22.4	12.4	2.74	40.6	20.7	5.30
	5	0.5	**29.2**	**2.3**	2.80	31.2	4.6	2.83	29.7	15.1	5.38
	5	-5.5	**31.9**	**15.3**	5.41	33.8	18.3	5.85	42.2	22.1	6.78

directions less significant. Overall, the proposed method globally outperforms MESSL and SRP-PHAT, except for DRR $= 0.5$ dB and SNR $= 5$ dB, for which SRP-PHAT performs the best.

One can see from Table I that the proposed method outperforms the RTF-CT-CGMM method for all acoustic conditions. Therefore, it is confirmed that the proposed CTF-based DP-RTF feature combined with the proposed consistency test provides more reliable features than the usual MTF-based RTF combined with the coherence test. As for Basic-CGMM, the DP-RTF estimation error for DRR $= -5.5$ dB will lead to noticeable spurious peaks, as was illustrated in Fig. 4. By suppressing the spurious peaks and/or strengthening the correct peaks, thanks to the entropy penalty, the proposed EP-MLE method achieves a significantly smaller outlier rate than Basic-CGMM, for a similar MAE. However, for DRR $= 0.5$ dB, there are much less spurious peaks, or they are much lower than the correct peaks. Thence, the proposed entropy penalty term is here less helpful compared with the low DRR case.

3) Blind Localization Results: Table II shows the blind localization results for the EP-MLE, Basic-CGMM and RTF-CT-CGMM methods. It can be seen that, for all three methods, the average of the MD rate and FA rate is generally larger than the outlier rate in the semi-blind configuration, which verifies that the blind configuration is more difficult than the semi-blind one. Also, for all methods and in a very general manner, both MD and FA increase when either the SNR or the DRR decreases, and when the number of speaker goes from two to three, which was expected. For the proposed EP-MLE method in particular, a larger DP-RTF estimation error is caused by more intense reverberations, which lead to more spurious peaks and peak shifts. For a given DRR, MD increases with the decrease of the SNR or with the increase of the number of speakers, since, as mentioned above, the method may suffer from a lack of sufficient number of DP-RTF observations. When the acoustic conditions get worse in terms of SNR or DRR, MAE increases due to the larger DP-RTF estimation error.

In general, MD, FA and MAE are considerably smaller for the proposed EP-MLE method (and for Basic-CGMM) compared to the RTF-CT-CGMM method, which is consistent with the results obtained for the semi-blind configuration. Unlike the semi-blind configuration, it can be seen that MD and FA are both smaller for EP-MLE than for Basic-CGMM, while the MAE are comparable, for almost all acoustic conditions (all except for MD at SNR $= 30$ dB, DRR $= 0.5$ dB, 2 speakers). This confirms the importance of the penalty term in the blind configuration. The semi-blind configuration inherently limits the FA score, and at the same time it can "force" the detection of low peaks, ensuring correct MD scores. In contrast, the setting of the threshold in the blind configuration favours either the MD or the FA. Therefore, in the blind configuration, it is more crucial to reduce the spurious peaks and enhance the correct peaks to facilitate the thresholding operation, which is exactly what is done by the entropy penalty term. By reducing the entropy to a proper extent, usually, the CGMM component weights corresponding to interfering directions are significantly decreased, while the weights of the true source directions are enhanced. As a result, MD and FA are both decreased by the entropy penalty term.

D. Results of Experiments with NAO Head Data

Table III and Table IV show the source localization results obtained with NAO head data, in the semi-blind and blind configuration, respectively. From Table III, it can be seen that

TABLE III: Semi-blind localization results for NAO data under various acoustic conditions. Here MC-MESSL denotes multichannel MESSL method. The lowest MD and FA among three blind methods for each acoustic condition are shown in **bold**.

	robot-to-source distance	EP-MLE (prop.)		Basic-CGMM		RTF-CT-CGMM		MC-MESSL [55]		SRP-PHAT [57]	
		Out(%)	MAE(°)	Out(%)	MAE(°)	Out(%)	MAE(°)	Out(%)	MAE(°)	Out(%)	MAE(°)
Two speakers	1.5 m	**8.5**	3.71	12.5	3.86	28.0	3.84	42.7	4.23	39.8	3.14
	2.5 m	**15.3**	4.93	21.0	5.20	24.5	5.81	44.8	4.65	36.3	4.68
Three speakers	1.5 m	**14.5**	5.21	17.3	4.46	34.7	4.66	46.1	4.77	44.2	3.58
	2.5 m	**18.7**	5.35	22.3	5.59	22.3	5.90	52.4	5.89	47.5	5.27

TABLE IV: Blind localization results for NAO data under various acoustic conditions. The lowest MD and FA among three blind methods for each acoustic condition are shown in **bold**.

	robot-to-source distance	EP-MLE (prop.)			Basic-CGMM			RTF-CT-CGMM		
		MD(%)	FA(%)	MAE(°)	MD(%)	FA(%)	MAE(°)	MD(%)	FA(%)	MAE(°)
Two speakers	1.5 m	**8.0**	14.3	3.79	15.5	**13.5**	3.80	33.5	22.0	4.36
	2.5 m	**12.8**	**18.0**	5.60	14.0	30.5	5.38	25.0	20.5	5.06
Three speakers	1.5 m	**17.8**	15.3	4.24	24.8	**15.2**	4.17	46.8	21.7	4.23
	2.5 m	**20.8**	17.7	5.37	21.8	24.3	5.45	37.2	**10.7**	5.23

the proposed method achieves the lowest outlier rate for all conditions, which verifies the effectiveness of the proposed entropy penalty and DP-RTF feature in such realistic scenarios. Overall, the multichannel MESSL method performs the worst. On the one hand, the spatially correlated noise and intense reverberation influence the performance as for the binaural case. On the other hand, the multichannel MESSL coordinates the microphone pairs through TF masks, rather than the usually used microphone calibration information. The time delay of each microphone pair is estimated using the results of source separation. This is advantageous for source separation that does not require information on microphone configuration. However, the known microphone configuration is necessary for source localization to specify the spatial relation between physical positions, namely to calibrate the time delays. In these experiments, the microphone calibration information is only used for source localization by comparing the calibrated time delays and the estimated time delays of MESSL, which leads to unsatisfactory localization performance. SRP-PHAT also has a high outlier rate due to the spatially correlated noise and real world reverberations.

In general, the performance measures reported in Table IV are consistent with the results obtained on the simulated data. Compared to Basic-CGMM, EP-MLE has smaller MD under all conditions, smaller FA under two conditions out of four (and for the other two conditions, the FA values for both methods are very close), and a comparable MAE. Also, the proposed method significantly outperforms the RTF-CT-CGMM method, since, again, the quantity and the quality of the observations are both higher for the proposed DP-RTF features than the RTF features based on the coherence test.

VI. Conclusion

In this paper, we presented a method for multiple-source localization in reverberant and noisy environments. The method is based on the model of [11] with the following original contributions: (i) the use of an entropy-based penalty term which enforces sparsity for the estimation of the model parameters, implemented via a convex-concave optimization procedure that is more efficient than an EM algorithm, (ii) the use of DP-RTF features, providing localization that is robust to both noise (thanks to the inter-frame power spectral subtraction) and reverberations, and (iii) the proposed consistency test algorithm that ensures that DP-RTF features are estimated from frame regions associated to a single active speaker, thus making possible to use these features for multiple-speaker localization. Overall, experiments conducted on both simulated and real-world data show that (i) the proposed DP-RTF features are more reliable than classical MTF-based features, for instance RTF features, (ii) the proposed CGMM model with DP-RTF features provides a better source localization compared to three baseline methods (RTF-based, MESSL, SRP-PHAT) in a semi-blind configuration, and (iii) the entropy penalty term used in the proposed localization technique makes it able to better localize the sources compared to the basic version of the same method (i.e. without the entropy penalty term); this is especially true in a blind configuration where the proposed method is efficient in *jointly counting and localizing the sources*. The experiments showed that the entropy-based penalty significantly improves the localization performance in terms of missed detections and false alarms.

In this study, the entropy-based penalty weighting coefficient γ was set to an empirical fixed value leading to good overall performance for all tested conditions. In future work, a principled setting of γ could be investigated, considering the noise level of the DP-RTF observations. Also, the DP-RTF features are more robust than MTF-based features at the cost of the need for more reliable data. An improved DP-RTF estimation process requiring less data will be investigated in the near future.

References

[1] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002, pp. I–529–I–532.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing,*, vol. 52, no. 7, pp. 1830–1847, 2004.

[3] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[4] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 68–77, 2010.

[5] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[6] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2233–2243, 2014.

[7] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[8] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and ℓ_1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[9] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing,*, vol. 58, no. 1, pp. 121–133, 2010.

[10] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[11] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1692–1703, 2015.

[12] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[13] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[14] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 320–324.

[15] X. Li, R. Horaud, L. Girin, and S. Gannot, "Local relative transfer function for sound source localization," in *The European Signal Processing Conference*, 2015, pp. 399–403.

[16] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.

[17] A. Deleforge and F. Forbes, "Rectified binaural ratio: A complex T-distributed feature for robust sound localization," *European Signal Processing Conference*, pp. 1257–1261, 2016.

[18] A. Deleforge, S. Gannot, and W. Kellermann, "Towards a generalization of relative transfer functions to more than one source," in *European Signal Processing Conference*, 2015, pp. 419–423.

[19] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[20] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

[21] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[22] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[23] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[24] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[25] X. Li, L. Girin, F. Badeig, and R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016, pp. 2819–2826.

[26] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[27] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[28] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[29] O. Schwartz, Y. Dorfan, E. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," in *International Workshop on Acoustic Signal Enhancement*, 2016.

[30] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1393–1407, 2016.

[31] ——, "Semi-supervised source localization on multiple-manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1477–1491, 2017.

[32] A. Deleforge, V. Drouard, L. Girin, and R. Horaud, "Mapping sounds onto images using binaural spectrograms," in *European Signal Processing Conference*, 2014, pp. 2470–2474.

[33] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound-source separation and localization on binaural manifolds," *International Journal of Neural Systems*, vol. 25, no. 1, 2015.

[34] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 4, pp. 718–731, 2015.

[35] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[36] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, 2007.

[37] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[38] J. Rousseau and K. Mengersen, "Asymptotic behaviour of the posterior distribution in overfitted mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 689–710, 2011.

[39] G. Malsiner-Walli, S. Frühwirth-Schnatter, and B. Grün, "Model-based clustering based on sparse finite Gaussian mixtures," *Statistics and Computing*, vol. 26, no. 1-2, pp. 303–324, 2016.

[40] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.

[41] H. Ishwaran, L. F. James, and J. Sun, "Bayesian model selection in finite mixtures by marginal density decompositions," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1316–1332, 2001.

[42] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[43] A. Asaei, M. Golbabaee, H. Bourlard, and V. Cevher, "Structured sparsity models for reverberant speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 620–633, 2014.

[44] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural computation*, vol. 15, no. 4, pp. 915–936, 2003.

[45] A. J. Smola, S. Vishwanathan, and T. Hofmann, "Kernel methods for missing variables," in *International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 325–332.

[46] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[47] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, pp. 263–287, 2016.

[48] A. Gilloire and M. Vetterli, "Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo can-

cellation," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1862–1875, 1992.
[49] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.
[50] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing*. Artech House Norwood, 2005, vol. 46.
[51] D. Campbell, "The roomsim user guide (v3. 3)," 2004.
[52] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR dummy-head microphone," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.
[53] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.
[54] H. W. Loellmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Proceedings of Speech Communication*, 2014, pp. 1–4.
[55] M. I. Mandel and J. P. Barker, "Multichannel spatial clustering for robust far-field automatic speech recognition in mismatched conditions," in *Proceedings of Interspeech*, 2016, pp. 1991–1995.
[56] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, M. S. Brandstein and D. Ward, Eds. Springer, 2001, pp. 157–180.
[57] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. I–121–I–124.

Xiaofei Li received the Ph.D. degree in Electronics from Peking University, in 2013. He is currently a Post-Doctoral researcher at INRIA (French Computer Science Research Institute), Montbonnot Saint-Martin, France. His research interests include multi-microphone speech processing for sound source localization, separation and dereverberation, single microphone signal processing for noise estimation, voice activity detection, and speech enhancement.

Laurent Girin received the M.Sc. and Ph.D. degrees in Signal Processing from the Institut National Polytechnique de Grenoble (INPG), Grenoble, France, in 1994 and 1997, respectively. In 1999, he joined the Ecole Nationale Supérieure d'Electronique et de Radioélectricité de Grenoble (ENSERG), as an Associate Professor. He is now a Professor at Phelma (Physics, Electronics, and Materials Department of Grenoble-INP), where he lectures signal processing theory and applications to audio. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It deals with speech and audio processing (analysis, modeling, coding, transformation, synthesis), with a special interest in multimodal speech processing (e.g. audiovisual, articulatory-acoustic, etc.) and speech/audio source separation. Prof. Girin is also a regular collaborator of INRIA (French Computer Science Research Institute), as an associate member of the Perception Team.

Radu Horaud received the B.Sc. degree in Electrical Engineering, the M.Sc. degree in Control Engineering, and the Ph.D. degree in Computer Science from the Institut National Polytechnique de Grenoble, France. Currently he holds a position of director of research with INRIA Grenoble Rhône-Alpes, where he is the founder and head of the PERCEPTION team. His research interests include computer vision, machine learning, audio signal processing, audiovisual analysis, and robotics. Radu Horaud and his collaborators received numerous best paper awards. He was an area editor of the *Elsevier Computer Vision and Image Understanding* (1999-2017), he is a member of the advisory board of the *Sage International Journal of Robotics Research* and an associate editor of the *Kluwer International Journal of Computer Vision*. He was program co-chair of IEEE ICCV'01 and of ACM ICMI'15. In 2013 Radu Horaud was awarded an ERC Advanced Grant for his project *Vision and Hearing in Action* (VHIA) and in 2017 he was awarded an ERC Proof of Concept Grant.

Sharon Gannot (S'92-M'01-SM'06) received his B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Israel in 1995 and 2000 respectively, all in Electrical Engineering. In 2001 he held a post-doctoral position at the department of Electrical Engineering (ESAT-SISTA) at K.U.Leuven, Belgium. From 2002 to 2003 he held a research and teaching position at the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. Currently, he is a Full Professor at the Faculty of Engineering, Bar-Ilan University, Israel, where he is heading the Speech and Signal Processing laboratory and the Signal Processing Track.

Prof. Gannot is the recipient of Bar-Ilan University outstanding lecturer award for 2010 and 2014. He is also a co-recipient of seven best paper awards.

Prof. Gannot has served as an Associate Editor of the EURASIP Journal of Advances in Signal Processing in 2003-2012, and as an Editor of several special issues on Multi-microphone Speech Processing of the same journal. He has also served as a guest editor of ELSEVIER Speech Communication and Signal Processing journals. Prof. Gannot has served as an Associate Editor of IEEE Transactions on Speech, Audio and Language Processing in 2009-2013. Currently, he is a Senior Area Chair of the same journal. He also serves as a reviewer of many IEEE journals and conferences.

Prof. Gannot is a member of the Audio and Acoustic Signal Processing (AASP) technical committee of the IEEE since Jan., 2010. Since Jan. 2017, he serves as the committee chair. He is also a member of the Technical and Steering committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the general co-chair of IWAENC held at Tel-Aviv, Israel in August 2010. Prof. Gannot has served as the general co-chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in October 2013. Prof. Gannot was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013 and EUSIPCO 2013. Prof. Gannot research interests include multi-microphone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation; dereverberation; single microphone speech enhancement and speaker localization and tracking.